DOCUMENT RESUME

ED 284 717 SE 047 593

AUTHOR Suydam, Marilyn N.
TITLE Evaluation in the Mathematics Classroom: From What and Why to How and Where. Revised Edition.
INSTITUTION ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Dec 86
CONTRACT 400-86-0016
NOTE 55p.; For first edition see ED 086 517.
AVAILABLE FROM SMEAC Information Reference Center, The Ohio State University, 1200 Chambers Rd., 3rd Floor, Columbus, OH 43212 ($8.50).
PUB TYPE Information Analyses - ERIC Information Analysis Products (071) -- Guides - Classroom Use - Guides (For Teachers) (052)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Attitude Measures; Cognitive Development; Criterion Referenced Tests; *Elementary School Mathematics; Elementary Secondary Education; *Evaluation Methods; Mathematics Education; *Mathematics Instruction; Norm Referenced Tests; *Secondary School Mathematics; Standardized Tests; Student Attitudes; *Test Construction

ABSTRACT

This document discusses the role and the scope of evaluation in the mathematics classroom. The scope of mathematics objectives to be evaluated, the scope of evaluation purposes in the mathematics classroom, and the scope of evaluation procedures are noted. Specific comments are made on various evaluation procedures, including: (1) observations; (2) interviews; (3) inventories and checklists; (4) attitude scales; (5) criterion-referenced tests; (6) norm-referenced tests; (7) standardized tests; and (8) diagnostic tests. Both general and specific suggestions for planning tests and for writing various test items are included. Types of test items discussed include multiple choice, true-false, matching, completion and essay. An extensive list of selected references is included to direct attention to documents which will provide additional help. (TW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED284717

MATHEMATICS EDUCATION REPORTS

Marilyn N. Suydam

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Robert N. Howe

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Evaluation in the Mathematics Classroom: From What and Why to How and Where

Revised Edition

ERIC Clearinghouse for
Science, Mathematics and Environmental Education
1200 Chambers Road - Third Floor
The Ohio State University
Columbus, Ohio 43212

December 1986

SE 047 593

OERI
Office of Educational
Research and Improvement
U.S. Department of Education

This publication was prepared pursuant to a contract with the Office of Educational Research and Improvement, U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions, however, do not necessarily represent the official views or opinions of the Office of Educational Research and Improvement.

Foreword

To some extent, this booklet is being revised at an inopportune time. Evaluation is the focus of attention of groups at both state and national levels, and much clarification and development is underway. Little of the results of current activities can be reflected at this point, but perhaps a revision will be needed sooner than the 12 years that have elapsed since this booklet was first published.

The impact of international, national, and state assessments of achievement on curricular goals is one cause for the focusing of attention. Data from the fourth National Assessment of Educational Progress in mathematics will appear in the near future, and it will surely be reviewed as carefully as the previous assessments have been. Information both on status -- how well are students achieving currently -- and on change -- what, if any, progress has been made since previous assessments -- is of vital interest. Data from the Second International Study of Mathematics recently attained headlines, with the ranking of the United States well below almost every other country on most of the achievement scales. Results from state assessments in mathematics, collated by Suydam (1984), indicated some areas of strength and many areas of weakness. The public, as well as educators, desires improvement.

Mathematics educators also have reached consensus on the need for change in both curriculum and instruction. The National Council of Teachers of Mathematics began this decade by publishing An Agenda for Action: Recommendations for School Mathematics of the 1980s. One of the eight recommendations focused on evaluation:

> *The success of mathematics programs and students learning must be evaluated by a wider range of measures than conventional testing.*

Noting that "many people use test scores as the sole index of the quality of mathematics programs or of the success of student achievement," the Council made a concerted plea for evaluation measures which assess the full range of the goals of mathematics program, including not only skills but also problem solving and problem-solving processes. More recently, a Task Force was appointed to study the role that testing and evaluation should play in mathematics programs, and ways of putting into practice evaluation strategies consistent with the goals and objectives of mathematics education.

Concurrently, the Mathematical Sciences Education Board has identified evaluation as one of its major strands of interest. The Board noted that:

Methods of evaluation -- especially standardized, paper-and-pencil, multiple-choice tests of 'basic skills' -- are being used across the country without sufficient reflection and are themselves obstacles to the teaching of new methods and higher-order thinking skills, as well as to the use of calculators and computers. The nation is in the grip of a 'testing mystique' which has led to the widespread use of such tests in spite of repeated warnings that several premises upon which their use is based are open to serious question.

Working cooperatively with the NCTM, the MSEB is developing recommended standards or criteria for excellence in school mathematics. One part of this effort will involve the development of "guidelines for redesigning tests and other assessment mechanisms so they are properly aligned with the curriculum and provide meaningful evaluation of student achievement." Questions about the validity of existing tests, including the degree to which they match what is being taught, the continued use of tests that inhibit or prohibit curricular change, and the misuses of assessment information have all been raised.

The impact of technology is clearly a part of the need for reform. Computers can deliver adaptive tests that can reduce the length of tests while preserving precision, and at the same time standardize administration and, of even more import, make results immediately available to the teacher. Tests which admit the use of calculators must clearly be developed: ten years ago, when they first became cost-feasible for classroom use, it was inconceivable that their existence could be ignored. Moreover, tests that fail to take into account such vital curricular strands as probability and statistics or problem-solving processes have survived past their time.

This may be an inopportune time for this booklet to appear; in another way, there is no inopportune time for such a booklet. Its purpose is to help classroom teachers extend their awareness of ways to evaluate and their skills in developing appropriate evaluative measures. It may help them prepare for the future.

It is intended as a quick reference guide rather than as an encyclopedia on evaluation of mathematics instruction. Its aim is to help teachers to review, to supplement, to develop questions about a process they use every day. The list of references should help them delve further into answers for their questions.

Two emphases are foremost:

(1) Evaluation means much more than paper-and-pencil tests.

(2) Each evaluation measure should be as good as possible.

Research in classrooms has indicated that teachers use many evaluation procedures. So, with awareness that change must continue, let's turn to the classroom and the ways teachers evaluate . . .

Evaluation

The following statement is an official NCTM position.
It was developed by the Professional Development and
Status Advisory Committee and adopted by the Board of Directors.

INCREASING demands for accountability have led state and provincial legislatures and agencies, school districts, professional organizations, and teacher education institutions to expand their efforts to evaluate teachers' knowledge and performance. Although such evaluations are often used as a basis for decisions about admission to teacher education programs, eligibility for certificates, or advancement in the profession, the most important purpose of evaluation is the personal and professional growth of the individual teacher and the improvement of teaching. Consequently, evaluation should be a cooperative process between the teacher and the evaluators.

Evaluation includes the identification of goals by the teacher and the evaluators, the collection of information, and a collaborative dialogue between the teacher and the evaluators to reformulate, redirect, and refine goals for the future. Goals for personal and professional growth may include some that are mandated by the state or province, district, teacher education institution, or individual school, but the teacher must be an active participant in identifying goals of a more specific nature.

Evaluation should not be limited to a single instrument—such as to paper-and-pencil testing of the students' or the teacher's knowledge alone, to checklists of isolated behaviors, or to a single observation session. Data should be gathered from various sources, including, but not necessarily limited to, the teacher, peers, students, supervisors, and administrators.

The use that is made of the information gained through the evaluation process is as important as the act of evaluation itself. The appropriate outcome of this ongoing process is a collaborative dialogue between the teacher and others involved in the process, resulting in a mutually agreed-on plan for professional growth.

Although the process of evaluating the effectiveness of teachers may be applicable across subjects and grade levels, the specific goals, criteria for observation, and resulting dialogue must be directly related to the content of mathematics and to the teaching strategies. The evaluation team should represent expertise and experience in mathematics and the teaching of mathematics as well as in evaluation.

Therefore, the NCTM recommends that supervisors and administrators work closely with teachers to assure that the evaluation process is used to enhance the professional development of the teacher and increase the effectiveness of mathematics teaching.

(March 1987)

Table of Contents

Page

Foreword i

NCTM Position Statement on Evaluation iv

I. Introduction 1

II. The scope of evaluation 3

A. The scope of mathematics objectives to be evaluated 3

B. The scope of evaluation purposes 6

C. The scope of evaluation procedures 7

1. Observations 7

2. Interviews 8

3. Inventories and checklists 10

4. Attitude scales 10

5. Criterion-referenced tests 12

6. Norm-referenced tests 13

7. Standardized tests 13

8. Diagnostic tests 14

III. Developing tests 15

A. Planning the test 15

B. Writing the test items: some general suggestions . . 17

C. Short-answer questions or completion items 18

D. Multiple-choice items 19

E. True-false items 22

F. Matching items 23

G. Essay items 24

Page

H. Some related points 26

1. Item pools 26

2. Item analysis 26

3. Two definitions 28

IV. Concluding comment 29

Selected References on Evaluation 30

Evaluation in the Mathematics Classroom:

From What and Why to How and Where

I. Introduction

Imagine a classroom. Perhaps it's your classroom.

Imagine 25 or 30 students in that classroom. Perhaps they're your students.

Imagine the students sitting at desks.

*

Imagine you see the students clear everything off the tops of the desks, except for a pencil.

What did the teacher say at the point the asterisk appeared?

Imagine the sound of the teacher's voice. Insert the words the teacher says in place of the asterisk. The words are: "Clear your desks. Take out a pencil. You are now going to have a test."

When we think of evaluation in the mathematics classroom, tests come immediately to mind . . . tests where students sit at desks and write or circle or draw lines.

But is that all there is?

Imagine that same classroom three days ago. Groups of students are scattered around the room. Two are spinning a three-colored cube, and making a record of what color lands upward each time. Several are making a graph on a bulletin board. Others are stretching yarn against various objects in the room. Some are seated with diagrams and worksheets, with games, with other materials before them.

Where is the teacher? What is he or she doing?

Is any evaluation occurring in the classroom at that moment?

Imagine the classroom four days ago. The students sit at their desks. The teacher stands near the chalkboard. She writes some numerals on the board. She asks a question. Several students in turn respond. She asks another question. One student comes to the board and draws a diagram. The teacher queries the group by raising her eyebrows. Three students shake their heads "no", four nod "yes", the others look puzzled. The teacher asks another question.

Is any evaluation occurring in the classroom during this lesson?

Imagine the classroom five days ago. The students have moved their desks so they have tables grouped by fours. Each group follows the directions of a leader as they manipulate materials on the desks. They help each other; they talk about what they find happening. Then each records a response on a worksheet.

Is any evaluation occurring as they work on this lesson?

The answer to each question is obvious. If the teacher is teaching, the teacher is evaluating almost every minute on each of the days imagined -- and on any other day you want to imagine. Sometimes the evaluation leads to an immediate reaction: you smile approval, or you frown; you say "good answer!", you say "that's on the right track"; you word a question so the student might see an error in the last response, you skip several questions because students are ready to move more quickly; you introduce a subtraction sentence instead of working only with objects, you get rods as an alternative way of clarifying a mathematical idea. Sometimes the evaluation leads to notes on students: anecdotal records, a comment on a problem to pursue further, a change of lesson plans for next week.

Evaluation in the mathematics classroom consists of much more than a testing program involving paper-and-pencil tests on mathematical content. Measurement of the content goals of mathematics is comparatively easy: you can readily obtain an objective measure of certain computational skills and specific mathematical processes that form a portion of the mathematics curriculum. Measuring other goals of the mathematics curriculum -- such as problem solving -- is more difficult. Evaluation includes a wide variety of means of collecting evidence on students' behaviors -- rating scales, questionnaires, checklists, reports from parents, student activities, and samples of students' work all provide useful evidence of behavior and progress. Observing, listening, presenting a task, interviewing: each makes a valid and viable contribution to the evaluation process.

But sometimes you evaluate with paper-and-pencil tests. Paper-and-pencil instruments have their place: they supplement other forms of evaluation. The very process of preparing for and taking a test helps students to synthesize what they have learned. The responses to specific items help the teacher to diagnose a weakness or confirm what was seen in the day-by-day process of observing student reactions and behaviors. Both students and teachers take stock: this mathematical idea or fact or skill or concept has been mastered and can be used in developing newer content. Another mathematical idea or fact or skill or concept needs to be given more thought or practice or development.

One of the purposes of this booklet is to help you to develop better paper-and-pencil measures. Tests continue to be a part of the educational environment, if only because they provide a feasible way of finding out, in a relatively short amount of time, what or how well

each child is learning certain content. Tests yield concrete and detailed evidence economically and in convenient form. Tests are, however, only tools whose value lies not merely in their use but in the skill and understanding of the teacher. Good tests do not just happen: they require much thought and careful planning and thorough analysis.

Another purpose of this booklet is to review other possible approaches to evaluation. What they are and how they can be useful are each considered. Finally, some pertinent literature on evaluation in mathematics is noted. Some references are inserted directly into the text; most are included in the list of references at the end of the booklet without being cited.

II. The scope of evaluation

Evaluation is a continuing, integral aspect of mathematics teaching and is essential for improving instruction. Evaluation ascertains whether the teacher is teaching what he or she thinks is being taught and the learner is learning what the teacher thinks the learner is learning. Thus, there must be a match between what is being taught and what is evaluated. Evaluation is qualitative as well as quantitative. It involves appraisal as well as measurement, for it includes the stage of making value judgments. This stage occurs when the means of evaluation is chosen, when it is applied, and when the results are judged.

Evaluation takes a variety of forms, since there is no one technique that is equally appropriate for measuring all aspects of learning. Both cognitive factors and affective factors must be assessed: the feeling and the doing aspects as well as the knowing and the thinking aspects are important in every mathematics program.

A. The scope of mathematics objectives to be evaluated

Scope-and-sequence charts in textbooks and curriculum guides provide one way of determining the dimensions of the mathematics program. Some mathematics educators have described the scope in various ways; for example:

> *In the study of mathematics a student must learn facts, develop concepts, use symbols, and master processes and procedures. But he [or she] should also learn to develop generalizations and to sense the presence of mathematical ideas and structures not only in abstract situations but also in many areas of human activity. He [or she] should develop his [or her] reasoning powers in order to prove or disprove a statement by deduction or to predict an event with appropriate probability. It is the function of*

evaluation to determine how well a student has mastered these varied aspects of mathematics.

[Sueltz, 1961]

Other writers have developed models to aid in the process of designing instructional materials and tests. The taxonomy developed by a committee working with Bloom has long provided a basic model for the analysis of educational goals in general (Bloom, 1956; Bloom et al., 1971; Krathwohl et al., 1964). Bloom's Taxonomy is presented in terms of two domains, the cognitive and the affective. The cognitive domain, not surprisingly, has been of most concern to those evaluating mathematics instruction, even though the importance of the affective domain is recognized. Goals in the cognitive domain have been organized into six main categories:

1. Knowledge -- recognizing or recalling specific material
2. Comprehension -- grasping the meaning of material
3. Application -- using information in concrete situations
4. Analysis -- breaking down material into its parts
5. Synthesis -- putting together parts to form a whole
6. Evaluation -- judging the value of material and methods for given purposes

Goals in the affective domain are organized into five categories: receiving, responding, valuing, organizing, and characterizing by a value. These categories have not been used at all as frequently as those in the cognitive domain.

Other models have been developed that are more specific to the goals of mathematics education. Generally, such models have combined some categories of Bloom's Taxonomy. Or they have used labels more specifically identified with mathematics. Thus, the School Mathematics Study Group (SMSG) used four categories to assess the cognitive domain: computation, comprehension, application, and analysis.

More recently, the National Assessment of Educational Progress (NAEP) has modified the framework used for planning the evaluation of mathematics objectives. For the fourth national assessment, the mathematics objectives are organized into five broad areas (NAEP, 1985):

1. Problem solving/Reasoning
2. Routine application

3. Understanding/Comprehension

4. Skill

5. Knowledge

Higher-order thinking skills, familiar applications, interpretations of underlying concepts and relationships, routine manipulations with standard procedures, and recall and recognition of mathematical content are thus assessed by the five categories.

These models have each aided curriculum developers and test constructers. Yet many teachers find it difficult to recall the categories, and even more difficult to apply them. Pikaart and Travers (1973) simplified the model so that it would really help teachers to describe specific learning goals, yet be comprehensive, flexible, and functional. They described three dimensions -- goals or products, content, and teacher behavior or processes, including planning, teaching, and evaluation. They noted that in practice it is difficult to distinguish activities that are planned for either cognitive goals or affective goals, since these are interrelated and interwoven in instruction. Therefore, the same model may be considered for both domains:

1. Knowledge

    a. Statements
    b. Basic skills

2. Understanding

    a. Concepts
    b. Principles

3. Problem solving

    a. Formulating hypotheses and testing them
    b. Proving theorems
    c. Solving non-routine problems

Categories or levels are important to consider in setting goals and developing objectives for instruction, in planning instructional activities and procedures, and in evaluating instructional outcomes. Too frequently, mathematics evaluation encompasses only the lowest level -- knowledge. It is easy to construct an objective test at the knowledge level; it is much more difficult to construct tests and other evaluation procedures that assess higher cognitive levels. Perhaps the greatest contribution of a model detailing the various categories is that it makes everyone aware of the need to evaluate higher-level outcomes.

B. The scope of evaluation purposes

Every teacher evaluates for at least three purposes:

1. *To assess the mathematics program in the classroom and in the school.*

The success of your mathematics program is not determined by how well it compares with the program in other schools. The important concern is the impact that it has on helping your particular students to learn mathematics. Is the content appropriate for your students? How well are they progressing toward the mathematical goals you have set? Are they able to apply their knowledge and skill in new directions? Does the program make the students want to continue to learn more mathematics? Do they enjoy doing and using mathematics? Is the mathematical content important and worthwhile? Is the program teachable and learnable?

Comparisons with other students in other schools can help you to attain some perspective on how well your students are doing, however. The National Assessment of Educational Progress (Carpenter et al., 1978, 1981; NAEP, 1983) and various state assessment programs (Suydam, 1984) attempt to provide such perspectives. But you are not teaching "other students in other schools". Your goal must be to help all of the students in your classroom to learn and to enjoy mathematics as well as each is able.

A guide to assessing the mathematics program in the school which reflects more than accountability test scores has been prepared by the National Council of Teachers of Mathematics (NCTM, 1981). It notes that, to provide a comprehensive mathematics program, the total school staff must be committed to:

1. meeting the needs, abilities, interest, and capabilities of each student.

2. exhibiting positive attitudes toward mathematics;

3. developing positive student attitudes toward mathematics;

4. preparing students to use mathematics successfully in their future vocations, avocations, and leisure time.

Twenty-one standards are then presented, concerned with instruction, the curriculum and instructional materials, the teacher of mathematics, and physical facilities and equipment. Appropriate questions to assess the attainment of each standard can be of help in evaluating the mathematics program.

2. To assess the achievement of the students in each classroom.

The vital factor to note in assessing achievement is that you must evaluate students in terms of both progress and status. Testing supplements other evaluation procedures as a means of ascertaining how well students have succeeded in mastering vital content and acquiring important skills.

3. To diagnose individual strengths and weaknesses.

You can use test results to place students in instructional materials, to group students for instruction, and to assign grades. You can also use test results to help you to learn more about how to teach more effectively.

Far too many mathematics tests consist simply of examples for which students are to provide answers. Far too often these tests are corrected by a check for correct and incorrect answers. The teacher who merely obtains the total score made by a student on a test is overlooking the greatest value of the test for instructional purposes. Alas -- so much is thrown in the wastebasket! Analysis of how the student reached the correct or incorrect answer is much more important than merely whether the answer was right or wrong. Analysis of performance on individual questions can tell you more than a total score can.

Evaluation procedures other than tests are invaluable in providing diagnositc information. As you listen and observe, you build the basis for interpreting test scores and deciding how to structure your teaching.

C. The scope of evaluation procedures

This section contains comments on various types of evaluation techniques: first, non-paper-and-pencil procedures; then, paper-and-pencil instruments.

1. Observations

Many mathematics lessons have a component in which students work in small groups or individually on tasks, assignments, or worksheets. This is a time when evaluating students' mathematical behavior is of singular importance. You can move about the room, observing students as they work, listening as they talk among themselves, making notes, questioning, making suggestions. You also observe during discussion periods, but your involvement in the discussion sometimes keeps you from attaining perspective: then you need to use your evaluation immediately as you continue the discussion. You have little chance to make notes. Your primary purpose is to guide. When you are free to observe as children work independently, you can evaluate even more effectively, with a defined perspective, and you can limit your observation to specific aspects of student behavior.

Note the method of attacking problems used by a student, and how he or she proceeds to work on a problem. Note the expression on Sue's face, her mannerisms, her concentration. Note how consistently she works, where she meets difficulty, when she becomes careless. Observe the emotional climate of the room. Observe the student's level of independence. Does Mark really need your help when he raises his hand, or does he need encouragement or praise? How dependent is he on help from you, from textbooks, from other students? Does he try various ways of solving a problem, or does he try to apply the last procedure used in class?

Make a simple memo that describes the situation and the behavior you've observed -- an anecdotal record. Use a small notepad or cards.

Name	*Date*	*Situation*	*Behavior*	*Comment*
Sue	1|17	group lesson, developing meaning of fractions with graph paper	quick to help neighboring students	
	1|20	computation game	missed most combinations in which she had to multiply by 7 or 9	redevelop and practice multiplication with 7 and 9

File the anecdotal records in the student's folder, in which you also place examples of daily work, project reports, and other papers.

Sometimes audiotape or videotape can be used to provide a record that you can go back over and analyze in more detail than when you are involved with the group. Photographs can provide a record of project work and "products". You can compare progress with more objectivity than simply through memory of what was done.

2. *Interviews*

An interview is an attempt to remove the restriction of writing, both that involved in your development of a test item and that of the student in developing an answer. You can delve more precisely into how a student solves an example or problem. You can learn how he or she goes about finding answers. You can follow as he or she describes what he or she is doing, and why.

Basically, the interview procedure is simple (Weaver, 1955):

(1) Face the student with a problem.

(2) Let him or her find a solution, as he or she tells you what he or she is doing.

(3) Challenge him or her to elicit his or her highest level of understanding.

Present Pat with an example written on a card:

46 $\overline{)327}$

Have him explain the procedure he follows while computing the answer.

Make notes as he works: sometimes it's helpful to have an exact record of what he says. Challenge him with such questions as, "Are you sure that's correct?" "What if I said the answer was ___?" "Is there any other way you could find the answer?" And remember that the two most important questions in an interview are "How?" and "Why?".

Other suggestions for interviewing include:

(1) Establish rapport and maintain a relaxed atmosphere. The student needs to understand what he or she is to do. You don't want Karen to search for the answer she thinks you want -- you want her answers, not yours. And you want to know what she's thinking. You want her to respond naturally, freely, and fully.

(2) Select your examples and questions for your purpose. At times, you'll interview only some students; at other times, the whole class. Use more than one example of a particular type, to determine how consistently a student works.

(3) Don't teach: don't give answers, and avoid leading questions and suggestions. Do as little talking as you can. You want to find out what the student is thinking.

(4) Record the student's answers and thinking and whatever he or she does, as you go. You may want to write fast, or tape record, or categorize or code, using an interview form. Don't rely on memory to make a "true" record after the interview is over. Careful records will enable you to ascertain patterns and provide other evidence for diagnostic teaching.

(5) Time may be a problem, or it may be an excuse. If you are serious about using interviewing as a means of finding out more about what students have learned and are learning, the time can be found -- when others have a worksheet or other seatwork, during free-reading time, etc. Schedule time one day a week or some time each day.

(6) You may want to have a student use a tape recorder without you being present. Have Kim tell how he does some aspect of

mathematics, why he attacks a problem as he does, why he likes or dislikes mathematics. A group of students might discuss various ways of solving a problem. You can play the tape back later and analyze student thinking more carefully and from a different perspective than you can if you're involved in the interview.

Researchers have often used interviews to assess the extent to which students understand a procedure or can apply a process. Thus, Lankford (1974) had seventh graders add, subtract, multiply, and divide with whole numbers and fractions. His compilation of students' responses can be of aid to you as a teacher, for the myriad errors that students make can help you plan instruction to avoid them or clarify meanings that are essential. Reys et al. (1982) used interviews to determine how good estimators work. Not only did this lead to the development of materials to teach estimation, but it also provides clues for you about how estimation skills are used.

3. Inventories and checklists

An inventory is a check of what the student knows about a specific topic or about the total program. It's probably especially useful at the beginning of the year. In oral form, primary-level teachers find it an indispensible alternative to a written test. At upper levels, it may be written and administered just as any other test is. The inventory frequently is used to survey the previous year's work or the status of students (both individuals and class) as they begin work in your classroom. Such a test is an aid in assessing the readiness of students for more advanced work, as well as a diagnostic aid. List the items and skills you want to inventory. Decide how you will inventory each: what directions will you give the students, what tasks and materials will you use, or what test items will you need.

A checklist is a type of inventory: a list of kinds of behavior to look for -- for example, evidence of interest in mathematics, applying mathematics, working with others, using a range of materials, etc. Rating scales are like checklists but provide for a degree of appraisal:

turns in assignments: never -- occasionally -- always

counts on fingers: frequently -- sometimes -- never

4. Attitude scales

Everyone believes that the affective component of learning is important: if students are interested in and enjoy mathematics, they'll learn it better. Attitudes involve both cognitive and non-cognitive aspects, an intellectual appreciation and emotional reactions. Thus, attitudes toward mathematics involve many facets, ranging from awareness of the structural beauty of mathematics and

of the important roles of mathematics, to feelings about the difficulty and challenge of learning mathematics, to interest in particular types of mathematics or particular methods of being taught mathematics.

Students' attitudes toward mathematics are assessed in several ways. One primary way is through observation: by observing expression, comments, and behaviors as a student reacts in a mathematical situation, you can infer how he or she feels about mathematics. You can note how often Jennifer chooses a mathematical activity when she has an option, how readily she attempts to apply mathematical ideas to real-life situations, how enthusiastically she reacts in a mathematics lesson. A checklist can be used as a systematic approach to recording observations.

At times, you can ask the student to comment directly on his or her attitudes. You can have Kai write an essay on a question such as, "Do you generally like or dislike mathematics? Why or why not?" Or she can be asked to complete sentences such as "I like mathematics because ---." You may ask her to rank in order of preference the subjects which she is studying: from this the level of her preference for mathematics can be inferred, by noting where she places it in relation to other subject areas.

Perhaps the most widely used measure of attitudes is the attitude scale. Half a dozen scales have been extensively used; on many of them, items such as the ones on the following scale appear.

Attitudes Toward Mathematics

(Scale Form B)

Marilyn N. Suydam and Cecil R. Trueblood
The Pennsylvania State University

This is to find out how you feel about mathematics. You are to read each statement carefully and decide how you feel about it. Then indicate your feeling on the answer sheet by marking:

A - if you strongly agree
B - if you agree
C - if your feeling is neutral
D - if you disagree
E - if you strongly disagree

1. Mathematics often makes me feel angry.
2. I usually feel happy when doing mathematics problems.
3. I think my mind works well when doing mathematics problems.
4. When I can't figure out a problem, I feel as though I am lost in a mass of words and numbers and can't find my way out.

5. I avoid mathematics because I am not very good with numbers.
6. Mathematics is an interesting subject.
7. My mind goes blank and I am unable to think clearly when working mathematics problems.
8. I feel sure of myself when doing mathematics.
9. I sometimes feel like running away from my mathematics problems.
10. When I hear the word mathematics, I have a feeling of dislike.
11. I am afraid of mathematics.
12. Mathematics is fun.
13. I like anything with numbers in it.
14. Mathematics problems often scare me.
15. I usually feel calm when doing mathematics problems.
16. I feel good toward mathematics.
17. Mathematics tests always seem difficult.
18. I think about mathematics problems outside of class and like to work them out.
19. Trying to work mathematics problems makes me nervous.
20. I have always liked mathematics.
21. I would rather do anything else than do mathematics.
22. Mathematics is easy for me.
23. I dread mathematics.
24. I feel especially capable when doing mathematics problems.
25. Mathematics class makes me look for ways of using mathematics to solve problems.
26. Time drags in a mathematics lesson.

This scale attempts to ascertain, less directly and therefore, it is hoped, with greater reliability or credibility, how strongly the student likes or dislikes mathematics. The major advantage of a scale such as this one is that it is designed to be used in a relatively short amount of time -- five to ten minutes. Its shortcoming is that it does not provide information across the range of factors that comprise attitudes toward mathematics. One of the most widely used scales of this multi-dimensional type is the one developed by Fennema and Sherman (1976). It assesses such facets as attitude toward success in mathematics; stereotyping mathematics as a male domain; the perceived attitudes of mother, father, and teacher toward one as a learner of mathematics; effectance motivation in mathematics; confidence in learning mathematics; and the usefulness of mathematics. The 26-item scale above has been used at all grade levels from kindergarten up, while the multidimensional scale is more appropriate for use with students who are in middle schools, secondary schools, or college.

5. Criterion-referenced tests

Paper-and-pencil instruments can help as you evaluate the individual student in terms of his or her own progress: what has Bob learned that he didn't know before you taught that unit on fractions or binomials? You compare the performance of a student with his or her previous performance. You design a test to ascertain whether or

not each student has learned what you have taught. You set a level that says, "if a student gets this percentage of the items correct, adequate mastery of the topic can be assumed." You can also ascertain how well your class has mastered a particular topic, so the test parallels the work in class. Such tests are criterion-referenced tests or mastery tests. Besides telling you how well a topic was learned, it also indicates the points at which you need to provide reteaching.

6. Norm-referenced tests

Paper-and-pencil instruments can also provide you with information on the status of the student in relation to other students in the class. A student is compared with others, with his or her achievement evaluated relative to the achievement of the class or a group of classes. The test may also be designed in terms of ascertaining whether students have been learning what you think they should be learning from your teaching. But instead of setting a mastery level, a scale is used: you expect a few students to do very well, a few to do poorly, but most to attain an "average" level. These tests are based on the content you have taught, as are criterion-referenced tests, but they are norm-referenced measures because the performance of the student is compared to that of the class.

7. Standardized tests

Another form of norm-referenced test is used in almost every classroom at least once a year: the commercially-published standardized test. While a few standardized tests are criterion-referenced, most are norm-referenced. Standardizing a test refers to developing prescribed, uniform requirements for administration and scoring and to the statistical analysis after the test is given to a large, specified group of students, resulting in the development of norms. These are expectancy levels: the scores that students in the norming poplulation attained. With the use of norms based on what students in many classrooms have scored, you have a measure of how well your students are learning when compared with many others.

Standardized tests are not a substitute for teacher-made tests, but a complement. More careful preparation and research are provided in developing a standardized test than is ordinarily possible for any individual teacher to provide when developing his or her own classroom tests. The content has been determined on the basis of common elements of widely used courses of study and textbooks. But standardized tests assess only a portion of the content that might be covered at a particular grade level or in a particular course. It is imperative that care be taken to ascertain that the standardized test adequately covers the expected outcomes of your school's mathematics program. Producers of reputable standardized tests publish outlines

of test content to compare with your local program. Aspects that are unique to your program will naturally not be included, so you'll have to make provision for assessing them.

Some guidelines have been suggested for selecting a standardized test:

(1) Formulate clearly the purposes that will be achieved by use of the test: precisely what kinds of information are the tests expected to supply? What outcomes are to be measured? What use is to be made of test results?

(2) What tests are available that will meet your needs? Lists of tests are available and should be consulted (Mitchell, 1985; NCTM, 1981).

(3) Obtain copies of those tests which, from their descriptions, appear to meet your purposes. Most test publishers will furnish sample test materials.

(4) Examine the tests and the test manuals for appropriateness for your particular needs, reliability, ease of administration and scoring, kinds of normative data provided, and evidences of careful development. Norms should have been established in schools similar to yours. There should be at least several thousand students in the norm group if the norms are to be accepted with confidence. The norm should be stated in a convenient form, such as percentiles (which indicate the percentage of students whose performance is found to be below any score) or grade norms (which show how well the average student in a specified grade has performed). The manual should include explicit directions for administration and suggestions for interpreting and using the results. Make sure that the time requirements are reasonable in terms of your school.

It seems safe to state that no students can avoid standardized tests as they progress through school. Therefore, it is wise to teach students how to take such tests: just reading the standardized test directions as they begin the first test is not enough. Develop tests that use the same types of items that will be met on standardized tests. This is particularly necessary for young children: many rarely see a multiple-choice item, for instance, until it is met on a standardized test.

8. Diagnostic tests

Some standardized tests are planned specifically to be diagnostic. They usually cover a limited scope in much greater detail than a test of general achievement. They are arranged to give scores on the separate parts.

You can also develop a teacher-made test that is diagnostic. The value of this type of test will depend on its ability to reveal specific weaknesses in the achievement of individual pupils. When you have identified the point at which the student begins to have difficulty, you can begin to help him or her to overcome the difficulty. Knowing that the student attained a score of thirty percent on a division test provides you with little guidance on how to improve your instruction; knowing that the student attained an incorrect answer to 673 - 4 tells you little more. But knowing that Nell's answer to that example was 16 remainder 3 tells you that perhaps she needs help in understanding the placement of the answer in the quotient, that perhaps she needs help with place value, that perhaps she does not understand the algorithm. It provides you with some information to follow up on.

In developing a diagnostic test, select the examples with care: they must be examples which readily allow errors of the types you predict. Have students show all of their work -- even when you use multiple-choice or other types of items.

III. Developing tests

Effective classroom tests can do more than assess student learning. They can zero in on what has been taught, and help clarify ideas for students. In this section some suggestions for developing tests will be considered. These suggestions have been drawn from many sources (e.g., Gronlund, 1968). An attempt has been made to be comprehensive, although sample items have not been included for all ideas. Some general procedures will be given first: these apply to the planning and development of all types of instruments. Then some specific suggestions to consider in developing various types of items will be presented.

A. Planning the test

A well-planned test must be designed to accomplish the purpose it is to serve. Have the kinds of information that you hope to get from the test clearly in mind.

1. List the objectives to be assessed by the test.

Consider: what have you taught? What mathematical content and ideas are really important for the students to have learned? Test objectives should correspond to instructional objectives; instructional objectives suggest the type of evaluation procedure and test item to use. Remember that some objectives are best measured by non-paper-and-pencil procedures.

The objectives will vary in scope and number depending on the type of test. For a mastery test, it may be that each objective toward which you taught will be assessed by several questions. For

an achievement test at the end of a longer period of time, you must be more selective in choosing only the major critical points, those which are important in the hierarchy or as a "base" for future learning.

2. *Decide on the types of items to be constructed.*

The type of item depends on the nature of the objective to be measured. Once you have determined that an objective can be measured adequately by a paper-and-pencil item, you need to decide what type of item to use. Some mathematical objectives are measured well by short-answer or completion items, or by multiple-choice items; a few objectives are best measured by true-false or matching items. Such objective-type items (so-called because they can be scored objectively, with independent scorers obtaining the same results) measure knowledge and comprehension levels efficiently. A relatively large field of content can be sampled, for objective-type items can be answered quickly and one test can contain many questions. This broad coverage helps provide a reliable instrument. For higher-level outcomes, consider essay tests (yes, even for mathematics!).

3. *Decide on the number of items to be written for each objective.*

There are no simple rules for determining the "right" number of items to use for measuring each objective. The content of a test should reflect the relative amount of emphasis each objective has received in the actual instruction: thus, the number of items will be in proportion to the amount of emphasis. The level of the items will be similarly related to the objectives. Take into consideration whether the interpretation of results will be in terms of each separate objective or the test as a whole. And of course consider the amount of time available for administration of the test.

To help ensure that the completed test will give each objective the desired coverage, develop an outline of specifications to serve as a guide for item construction.

content (objectives)	*% of emphasis in instruction*	*number of items*	*level of items* K	U	upper
forming equivalent classes	10	4	1	2	1
adding 'like' fractions	20	8	2	3	3

Tests should measure an adequate sample of the learning outcomes and content included in the instruction. You can never ask all of the questions you might like to: you can only test a sampling of the most important outcomes.

B. Writing the test items: some general suggestions

The role of each item is to ascertain whether a student has attained the objective or not. There should be nothing about the structure or presentation of an item that leads those who know the correct answer to get the item wrong or those who do not know the answer to get the item right.

1. Select the measurement technique that is most effective for the specific objective.

2. Use clear, simple statements. Use language that students understand. Choose concise vocabulary, and sentence construction that is appropriate to the level of your students. Break a complex sentence into two or more separate sentences.

3. Design each item so that it provides evidence that an objective has been achieved. Avoid testing for unimportant details, unrelated bits of information, or irrelevant material.

4. Check items against the table of specifications to make sure that you have the desired emphasis on various content objectives at various levels of difficulty.

5. Work with another teacher or group of teachers in reviewing each others' items. Cut out points of doubtful importance or correct unclear wording.

6. Adopt the level of difficulty of a test item to the group and to the purpose for which it is to be used.

7. Initially, you may want to write more items than you will need on the final form of the test. Then you can discard weaker items. Many teachers write down items each day for possible inclusion on a test, to help ensure that important points will not be omitted.

8. Have each student work from a separate copy of the test, rather than from a test written on the chalkboard.

9. Number all items consecutively from the first item on the test to the last.

10. Avoid putting part of an item on the bottom of one page and the rest on the top of the next page.

11. If the form of a test or a group of items is unfamiliar, use sample items to help clarify the directions. Spend some time teaching students how to take a test.

12. Precede each group of items with a simple, clear statement telling how and where the students are to indicate their answers.

13. When you want students to show their work, provide adequate space near each item. "Boxing in" this space helps you to locate it quickly.

14. Begin a test with easy items. Placing difficult items at the beginning of a test is likely to discourage average and below-average achievers. You can then arrange items so that the test gets increasingly more difficult, or you can mix easy and difficult items.

15. Many times you'll need to have more than one type of item on a test (short-answer, multiple-choice, etc.). Place all items of one kind together. Always have more than one or two items of a particular type (except possibly of the essay type).

16. Avoid a regular sequence in the pattern of responses: students are likely to answer correctly without considering the content of the item at all.

17. Eliminate irrelevant clues and unnecessary or non-functional clues, but provide a reasonable basis for responding.

18. Make directions to the student clear, concise, and complete. Instructions should be so clear that students know what they are expected to do, although they may be unable to do it.

19. Prepare a key containing all the answers that are to be given credit. Make it so that it can be placed beside the answer spaces used by the students.

20. After the test, go over questions with your students: they can point out ambiguities and other errors, helping you to improve items for future use.

21. Analyze student responses to each item, for diagnostic use.

C. <u>Short-answer questions or completion items</u>

The short-answer item employs a question, an incomplete statement, or a computational example to elicit from the student appropriate words, symbols, or numbers. It is generally limited to questions that call for facts: who, what, when, where, how many. Many classroom mathematics tests are solely of this type: it is

frequently used to measure the ability to compute. You can present a number of computational exercises, or you can focus the student's attention on particular aspects of a computation.

In the completion item, certain important words or phrases are replaced by blanks to be filled in by the student. It must be very carefully prepared, or it is likely to measure only rote memory, or intelligence rather than achievement.

1. State the item so that only a single brief answer is required and possible.

2. Use a direct question when possible; switch to an incomplete statement only when greater conciseness is possible.

3. Words to be supplied should relate to the main point of the statement.

4. Blanks should be placed at the end of the completion statement.

5. Avoid giving extraneous clues to the answer.

6. If the answer can appear in more than one form, give specific directions about which form to use. Indicate such things as the degree of precision for numerical answers and whether labels must be used.

7. Avoid the use of sentences taken directly from the textbook. They are frequently ambiguous out of context, and encourage rote memorization.

8. Do not give clues to the answer by varying the number or length of the blanks.

D. Multiple-choice items

The multiple-choice item consists of a stem which is a question or an incomplete sentence presenting a problem situation, followed by several alternatives, which are possible solutions to the problem. One of the alternatives is the correct answer; the others are plausible answers, called distracters because their function is to distract students who are uncertain of the correct answer. The stem may also be a problem, graph, or diagram followed by the alternatives relating to it.

The ease of scoring undoubtedly plays a big part in the popularity of multiple-choice items. Student answers are easy to read and unambiguous. The use of computer-scoring has made the multiple-choice item virtually the only type used when a computer is available or for standardized tests. In general, scores on

multiple-choice tests are comparable to those that would be obtained from free-response tests, for the same level of content.

But there are other reasons for deciding to use multiple-choice items: they tend to provide a more adequate measure of many objectives than do other objective-type items. Multiple-choice tests have high reliability compared with other types of tests. And with careful analysis and development, the multiple-choice item can be adapted to most types of content and to most levels of objectives. It can assess the student's ability to recognize facts or relationships, to discriminate, to interpret, to analyze, to make inferences, to solve problems. Its biggest weakness is that it allows the student to guess, but this affects scores less than on other types of items.

Multiple-choice items should not be used when a simple question is adequate, that is, where there is clearly only one correct answer and no plausible distracters. They should not be used when there are only two plausible responses; a true-false item is usually effective in that instance.

1. Make directions explicit, so that students know exactly what type of response is required. Is more than one answer possible? Are they to select "the correct answer" or "the best answer"? How should they record answers? Shoud they guess if they aren't sure of the correct answer?

2. The stem should present a single worthwhile problem to be solved, expressed clearly and without ambiguity. State the question so there can be only one interpretation. Check on the clarity of the stem by covering the alternatives and determining whether the question could be answered without the choices.

3. Make each question independent of other questions. Students are often able to select the correct answer to one item because of information gleaned from another item. Where an answer to one item is used in succeeding items, students who miss that item will miss the succeeding items.

4. Make alternative choices as brief as possible. Instead of repeating words in each alternative, include them in the stem.

5. State the stem in positive form whenever possible. When negative wording is used, emphasize it by underlining or by capitalizing.

6. The best alternative choices to the correct answer are those using commonly mistaken ideas or common misconceptions or errors commonly made by the students. Excellent distracters can be obtained from incorrect responses on short-answer, completion, or essay tests.

7. In general, use the same number of alternatives for each item on a test. But remember that an item is not improved by adding an obviously wrong answer merely to obtain another alternative. Generally four or five alternatives are used, to reduce the chance of guessing the correct answer. It is better to have only four alternatives when five plausible choices are not available.

8. Make all incorrect responses equally plausible or "attractive" to the student who does not know the correct answer. If plausible distracters are difficult to find, use another type of item rather than ineffective alternatives. The more homogeneous the alternatives, the more difficult the item will be. The correct answer is one which cannot be refuted.

9. Make all alternatives grammatically consistent with the stem, and parallel in form. Avoid verbal clues which might enable students to select the correct answer or to eliminate an incorrect alternative: similarity of wording in the stem and the correct answer, for instance, or including two responses that are all-inclusive or two that have the same meaning. Check the structure by reading each alternative with the stem.

10. Do not consistently make the correct response longer or shorter than the distracters. There is a tendency to include the greatest amount of detail in the correct answer.

11. Avoid the use of qualifying words such as "always", "never", or "all" as much as possible: they are clues to a test-wise student that an alternative probably is not true.

12. Avoid use of the alternative "all of the above" and use "none of the above" with care. The inclusion of "all of the above" makes it possible to answer the item on the basis of partial information: the student can realize that it is the correct choice by noting that two of the alternatives are correct, or that it is not the correct choice by noting that at least one of the alternatives is incorrect. The chance of guessing the correct answer is thus increased. The use of "none of the above" may be measuring only the ability to detect incorrect answers: a student may do this and still not know the correct answer. If you want to reduce the chances of students estimating the answer without doing an entire computation (when that is the objective), use a completion-type item.

13. Avoid using a pattern for the position of the correct response. Students are quick to perceive patterns or apparent patterns and select their answers accordingly. Use some system of random order for the positions of the correct answers on each multiple-choice test -- and check to make sure that patterns did not inadvertently occur. Many teachers fail to use a, d, and e as

often as they use b and c as distracters. Students learn that their chances of guessing the correct answer are better if they guess b or c. Be sure the correct response is placed in all positions approximately the same number (but not exactly the same number) of times.

14. Control the difficulty of the item either by varying the problem in the stem or by changing the alternatives.

15. Use an efficient item format.

    a. List alternatives on separate lines, one under the other, making them easy to read and compare.

    b. Use letters in front of alternatives, to avoid confusion with numerical answers. For algebra tests, you might use numerals in parentheses.

E. True-false items

The true-false item can be difficult to construct, for statements must be unquestionably true or false. To construct such items to measure important outcomes is difficult: they adapt best to the measurement of specific facts, understanding of principles or generalizations, and common misconceptions. They can be used only when there are only two possible alternatives. Because they are highly subject to guessing, true-false items have little value as diagnostic tools.

"Alternative-response items" are variations in which the student must respond "agree" or "disagree"; "right", "partly right", or "wrong"; or with similar words. Other variations include items in which attention is directed to an underlined word or phrase; after deciding that any statement is false, the true words are to be inserted in place of the underlined words. Students can also be asked to state why the statement is true or false. Cluster true-false items deal with a single idea; such mathematical content as graphing can be tested with such an item, where students are asked to look at a graph and then respond to a series of true-false items about the data portrayed.

1. Have students circle T and F, or write T and F or + and 0 (rather than t and f or + and −, which cannot be distinguished as readily).

2. State the item clearly and specifically so that it is unequivocably true or false. Avoid the use, however, of specific qualifiers such as "always" or "never" -- or use them in both true and false statements. Check for ambiguities.

3. The item should deal with a single definite idea. The use of several ideas in each statement tends to be confusing and the item is more likely to measure reading ability than achievement. There should be no more than one problem-setting clause.

4. Avoid making true statements longer than false statements.

5. Make the crucial element readily apparent to the student. It is better to have the crucial element come at the end rather than in the early part of a two-part statement.

6. Have an approximately equal (but not exactly equal) number of true and false statements (vary the proportions from test to test).

7. Randomly arrange true and false items; check to be sure there is no inadvertent pattern.

8. Avoid trick statements which appear to be true but are really false because of some inconspicuous or trivial word or phrase.

9. Avoid statements that are partly true and partly false.

10. Avoid the use of statements extracted from textbooks. Out of context, such statements are often unclear or ambiguous.

F. Matching items

The matching item measures ability to discriminate between several items of similar material as they are related in a given way with items of another set. The matching exercise is essentially a modification of the multiple-choice form. When all of the responses in a series of multiple-choice items are the same, the matching format is more appropriate. Said another way, unless all of the responses in a matching item serve as plausible alternatives for each premise, the matching format is inappropriate.

Matching items can be used for such content as definitions and words defined, measurement and formulas, or geometric shapes and names. They are most appropriate for testing at the knowledge level; it is difficult to adapt them to testing for comprehension and higher-level goals.

1. Place the premise column on the left, the briefer responses on the right. Each of the items in the left column should have a test item number; the responses should be preceded by letters. Have students place answers to each item in a space to the left of the item number.

2. The items in the two columns must be homogeneous (that is, no responses should be logically excludable as answers by a student

who is uninformed). If they are not homogeneous, students may be provided with clues which will help them to match the terms, resulting in easier test items. Selection of the correct match should be dependent on knowledge of the correct answer, not on ability to eliminate incorrect answers on the basis of extraneous information.

3. To reduce the effect of guessing, one column should contain more terms than the other. Directions should clearly indicate whether responses may be used once, more than once, or not at all.

4. Do not include too many items in either column: a maximum of twelve items in the premise column should be considered. Longer lists require too much searching time.

5. Place the items in the response column in some logical order, to enable students to scan the list quickly to find the term they had in mind. Jumbling the terms merely increases searching time, without increasing the probability of correct answers being located.

6. Be sure that there is only one response which is the correct match for each premise when responses are to be used only once.

G. Essay items

Essay items are not often used on mathematics tests, but they can and should be. Such items require students to do more than compute a solution or recall specific facts. They must think about mathematics and meaning. They must organize their own ideas and express themselves effectively in their own words, using both knowledge and reasoning. Purely factual information is not assessed as efficiently as with objective-type items, but higher levels of reasoning can be tapped. Essay questions can be used to assess comprehension, applications, and analysis outcomes; they provide a means of assessing a student's ability to synthesize or to evaluate mathematical ideas which is rarely provided by objective items. Essay questions that assess complex achievement are apt to include such key words as why, explain, compare, relate, interpret, criticize, develop, derive, classify, illustrate, and apply. Clearly, they assess higher-order thinking skills.

There are difficulties in using essay items, as you're aware. An essay test covers a limited field; the questions take so long to answer that relatively few can be answered in a given period of time. A representative sampling of content is not feasible. Essay items are subjective, more difficult to score, and less reliable than objective-type items. Scores are apt to be distorted by writing ability and by bluffing. Students who are fluent can often avoid discussing points of which they are unsure. But there are things you

can do to minimize these problems, beginning with the writing of clearly defined items -- general enough to offer some leeway, but specific enough to set limits.

1. Use essay questions to evaluate achievement on those objectives not readily tested by other types of items.

2. Phrase the questions as precisely as possible and be specific in wording, so the objective of the item is clear and students are made aware of the specific scope or limits to be included in the answer.

3. Make clear to students the basis on which the answer will be judged, such as content, organization, comprehensiveness, relevance, appropriateness, etc.

4. Require all students to answer all questions, so they are all taking the same test. One way of doing this is by setting time limits for each item. Be sure that students have time to write adequate answers: time must be allowed for thinking as well as for writing. Provide adequate space for answers (or have students write on separate paper).

5. Discuss ways of answering essay questions with the students.

Since scoring essay items can be difficult, here are some suggestions which will increase objectivity.

1. List specific objectives for each essay question as you write it. Evaluate in terms of the objectives. Separate scores may be given for style of writing or spelling, but should not "contaminate" the evaluation of the mathematical objective being assessed.

2. Write out the essentials of a complete answer to each question or prepare a model answer ahead of time. Use it in the same way in scoring each paper. This does not preclude adding other acceptable points made by students. Determine the number of points to be assigned to each part of the model answer, or determine criteria for levels of expected quality.

3. Keep the identity of students unknown. Have students use a coded numeral on the papers or have them write their names on the back or at the end of the test.

4. Read one question through the entire set of papers, scoring each item for all papers before going on to the next item.

5. More uniform standards can be applied by reading the answers twice. At the first reading, sort the papers into several piles. Then reread to check on the uniformity of answers in each pile

and make any necessary changes in rating. Assign the same item score to all papers in a pile.

6. Reshuffle the papers so that a paper may not be scored unduly high or low because of its position, after scoring each item.

H. Some related points

1. Item pools

An item pool is simply a collection of test items that you can put together in various combinations to form a test. Several items may be developed for testing each specific objective; you can select the one that best meets test requirements. You'll probably find that a card file is the easiest way of filing the items. Write each item on a card, noting the topic or objective in one corner. At the bottom or on the back, record what you've learned about the item: When it should be used, what percentage of students get it correct each time you use it, and so on.

Other sources of models for items include commercial tests, textbooks for students or teachers, collections of items or item banks, and the tests which were constructed for various research studies.

Item sampling is a technique for assessing the status of a group of students. The National Assessment of Educational Progress uses this procedure, in order to avoid having students take a lengthy test. Instead of having all students at age nine answer all items, many similar samples of students are selected and each answers varied portions of the items. Then the scores are combined to depict how well nine-year-olds, as a group, answered the questions. Since your focus is usually on how well students are achieving, rather than on how well content is being achieved across students, you will probably not use item sampling techniques. You may find the term appearing frequently in various articles about testing, however.

2. Item analysis

Item analysis is the process of studying the students' responses to each item. An item analysis can tell you how difficult an item is and how well each question discriminates between high- and low-ranking students. It's especially important if you are going to re-use the item: it can indicate whether or not an item needs to be revised. It's also useful even if you don't plan to use the item again, for it can tell you what questions are especially appropriate to test certain objectives. Or it can be used simply as part of your diagnostic procedures.

Computer programs are used for item analysis for tests that are developed for research studies, for standardized tests, and for other

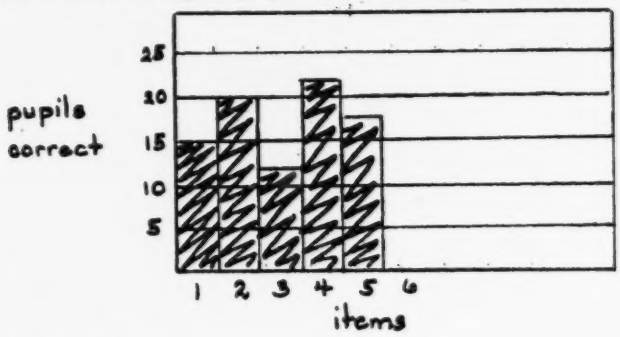
tests that will be used by many groups of students. Perhaps you have available a microcomputer program that can perform an item analysis. Unless you prefer to use it, however, only simple item analysis procedures seem warranted for most classroom tests. Here are several suggestions:

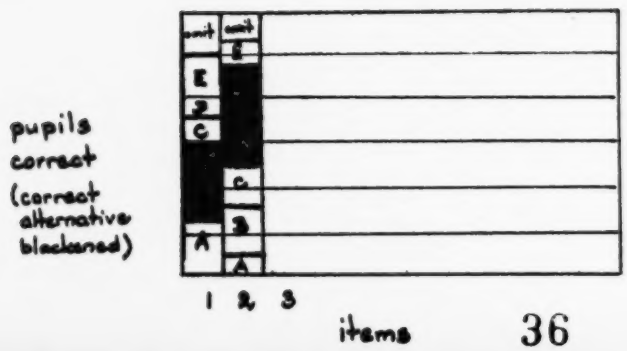
(1) Look at the test: what items were missed by many students? Were they missed because of a "fault" in the item or was there a "fault" in the instruction? What do you do next? Revise the item or revise the instruction.

(2) A simple measure of difficulty is the percentage of students who got the item correct. This gives you an approximation of how difficult the item is. By recording this information for each item in your item pool, you can build a test which will be at an appropriate difficulty level. This is especially helpful when you're developing a test in which you want to rank students; each item should then be of medium difficulty -- approximately 40% to 60%. (For mastery tests, your standards will be different.)

You can check the students' papers yourself to obtain the percentages, or you can do an item analysis by a show of hands. Call out the item numbers one by one and have students who have the item correct hold up their hands. Count and record the number of hands. Have students convert it to a percent, or do this yourself.

You can extend this activity by building a graph with the students, recording either the number of students who got the item correct or the number of incorrect responses. (For multiple-choice items, keep a record of the number selecting each alternative.) Are there any patterns in the graph? What items were missed most? Are there areas involving any particular objective?





(3) To do a more sophisticated item analysis, use this procedure:

(a) Arrange the test papers in order from highest to lowest score.

(b) Select the highest one-third and the lowest one-third (approximately), setting aside the middle one-third of the papers.

(c) For each item, count the number of students in the upper group who got it right and the number in the lower group who got it right. Let's say you have 10 papers in the upper one-third and 10 in the lower one-third. For one item, here's the count for the correct answer: *upper -- 7*
lower -- 3

(d) Convert these numbers to percentages:

$$\textit{for all students: } \frac{7+3}{20} = 50\%$$

$$\textit{for upper-ranking students: } \frac{7}{10} = 70\%$$

$$\textit{for lower-ranking students: } \frac{3}{10} = 30\%$$

Most items on a test used to rank students should be of medium difficulty, so this item appears to be at a satisfactory level of 50%. The harder the item, the lower the percentage of students getting it correct. Moreover, if the item is a good one for ranking students, then substantially more students in the upper group will have answered it correctly -- as happened in this case. Items on which many more students in the lower group got the item correct need revision. Thus, if the percentages above had been reversed, with 70% of the lower-ranking students getting it correct while only 30% of the upper-ranking students got it correct, there is something wrong with the item and it should be revised -- or discarded.

(4) On multiple-choice tests, determine the effectiveness of distracters by comparing the number of students in upper and lower groups who selected each incorrect alternative. A good distracter will attract more students from the lower group than from the upper group. Each distracter should attract some students or it is not serving effectively as a distracter. (Different criteria, however, apply to mastery tests.)

3. Two definitions

Any test, whether constructed by an individual teacher or commercially published, should meet several criteria, including acceptable validity and reliability. Validity pertains to the relevance of the test. Are you collecting the right kinds of information? Does the test measure the skills, understanding, or

knowledge that it was intended to test? Does it measure the significant behaviors that it was intended to test? Does it measure the significant behaviors that are specified in the objectives? Are all items relevant to those behaviors? Is the test a balanced sampling of the behaviors? Reliability pertains to the consistency of the test. How accurate and stable is the test? Does it measure the same achievement consistently? The nature of mathematics helps to make mathematics tests quite reliable. If a test were perfectly reliable, the students would have the same score or be ranked in the same order if the test were repeated, or a parallel form of the same test were administered. Reliability is commonly reported by a coefficient or correlation between forms of the test or between two halves of the same test. Perfect reliability is represented by a coefficient of 1.00. Usually a coefficient of at least .80 is expected on an objective mathematics test; many mathematics tests have reliabilities of .90 and higher. Tests of computational ability are usually more reliable than tests of problem-solving ability.

You probably have many other questions. Answers to these questions, whether about definitions or about testing or about other aspects of evaluation, may be answered by one or more of the references included at the end of this booklet. These references are grouped by major theme, to aid you in locating pertinent information.

IV. Concluding comment

The goal of evaluation is improving instruction. Measuring or assessing or testing only indicates: the teacher then has to do something as a result of what he or she has learned. This booklet has not attempted to consider the most difficult task in teaching: the use of the knowledge and understanding gained from evaluation. Evaluation is only a beginning . . . you must continue the process of teaching.

Selected References on Evaluation

These references, selected for their potential interest to teachers, are categorized under the following themes:

1. *Attitudes*
2. *Competency Testing*
3. *Criterion-Referenced Tests*
4. *Diagnostic Testing*
5. *General*
6. *Interviews*
7. *Item Pools*
8. *Materials Evaluation*
9. *Other Evaluation Techniques*
10. *Problem-Solving Tests*
11. *Program Evaluation*
12. *Special Populations*
13. *Standardized Tests*
14. *Teacher-Made Tests*

Attitudes

Fennell, Francis. Affective Assessment Strategies in a Diagnostic Prescriptive (Clinical) Mathematics Setting. April 1979. ERIC: ED 191 684.

Fennema, E. and Sherman, J. Fennema-Sherman Mathematics Attitudes Scales. JSAS Catalog of Selected Documents in Psychology 6: 31; 1976. (Ms. No. 1225)

Hodges, Helene L. B. Learning Styles: Rx for Mathophobia. Arithmetic Teacher 30: 17-20; March 1983.

Michaels, Linda A. and Forsyth, Robert A. Measuring Attitudes Toward Mathematics? Some Questions to Consider. Arithmetic Teacher 26: 22-25; December 1978.

Competency Testing

Carpenter, Thomas; Coburn, Terrence G.; Reys, Robert E.; and Wilson, James W. Results from the First Mathematics Assessment of the National Assessment of Educational Progress. Reston, Virginia: National Council of Teachers of Mathematics, 1978.

Carpenter, Thomas P.; Corbitt, Mary Kay; Kepner, Henry S., Jr.; Linquist, Mary Montgomery; and Reys, Robert E. Results from the Second Mathematics Assessment of the National Assessment of Educational Progress. Reston, Virginia: National Council of Teachers of Mathematics, 1981.

Carter, Betsy Y. and Leinwand, Steven J. Calculators and Connecticut's Eighth-Grade Mastery Test. Arithmetic Teacher 34: 55-56; February 1987.

Crosswhite, F. Joe; Dossey, John A.; Swafford, Jane O.; McKnight, Curtis C.; and Cooney, Thomas J. Second International Mathematics Study Summary Report for the United States. Champaign, Illinois: Stipes Publishing Company, 1985.

Hagan, Ronald D. Factors Influencing Arithmetic Performance on the Tennessee State-Mandated Eighth Grade Basic Skills Test. School Science and Mathematics 82: 490-505; October 1982.

Henderson, George L. and others. Wisconsin Mathematics Test, Grades 7 and 8. Madison: Wisconsin State Department of Public Instruction, 1978. ERIC: ED 051 185, ED 151 186, ED 069 475.

Mappus, L. Lynne and others. Mathematics: Teaching and Testing Our Basic Skills Objectives -- Grades 1, 2, 3; Grades 4, 5, 6; Grades 7, 8. 1981. ERIC: ED 226 056, ED 226 057, ED 226 058.

Mott, Warren. Implementing Mathematics Proficiency Testing. Mathematics Teacher 73: 19-22; January, 1980.

National Assessment of Educational Progress. The Third National Mathematics Assessment: Results, Trends and Issues. Denver: NAEP, 1983.

Ortiz-Franco, Luis. Patterns of Mathematics Minimum Competency Skills in the Elementary School. Los Alamitos, California: Southwest Regional Laboratory for Educational Research and Development, August, 1979. ERIC: ED 204 114.

Smith, William D. Minimal Competencies: A Position Paper. Arithmetic Teacher 26: 25-26; November 1978.

Suydam, Marilyn N. Assessing Achievement Across the States: Mathematical Strengths and Weaknesses. Columbus, Ohio; ERIC Clearinghouse for Science, Mathematics, and Environmental Education, December 1984. ERIC: ED 255 356.

Criterion-Referenced Tests

Besel, Ronald. Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests. February 1973. ERIC: ED 076 658.

Heines, Jesse M. An Examination of the Literature on Criterion-Referenced and Computer-Assisted Testing. November 1975. ERIC: ED 116 633.

Knipe, Walter H. and Krahmer, Edward F. An Application of Criterion Referenced Testing. February 1973. ERIC: ED 074 154.

Porter, Deborah Elena. Criterion Referenced Testing: A Bibliography. TM Report 53. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, December 1975. ERIC: ED 117 195.

Roudabush, Glenn E. and Green, Donald Ross. Some Reliability Problems in a Criterion-Referenced Test. February 1971. ERIC: ED 050 144.

Winkles, Jim. Criterion-Referenced Testing and Core Curriculum. Australian Mathematics Teacher 37: 8-11; August 1981.

Diagnostic Testing

Algozzine, Bob and McGraw, Karen. Diagnostic Testing in Mathematics: An Extension of the PIAT? Minneapolis: University of Minnesota Institute for Research on Learning Disabilities, March 1979. ERIC: ED 185 749.

Burns, Paul C. Analytical Testing and Follow-up Exercises in Elementary School Mathematics. School Science and Mathematics 65: 34-38; January 1965.

Dunlap, William P. and Brennen, Alison H. Blueprint for the Diagnosis of Difficulties with Cardinality. Journal of Learning Disabilities 14: 12-14; January 1981.

Engelhardt, Jon M. and Wiebe, James H. Measuring Diagnostic/Remedial Competence in Teaching Elementary School Mathematics. 1978. ERIC: ED 177 018.

Herman, Joan and Winters, Lynn. Test Design Manual: Guidelines for Developing Diagnostic Tests. Los Angeles: California University of Los Angeles, 1985. ERIC: ED 266 159.

Liedtke, Werner. Learning Difficulties: Helping Young Children with Mathematics -- Subtraction. Arithmetic Teacher 30: 21-23; December 1982.

McAloon, Ann. Using Questions to Diagnose and Remediate. Arithmetic Teacher 27: 44-48; November 1979.

General

Australian Council for Educational Research. Mathematics Evaluation Procedures K-2, for Use by Teachers with Children in Years K-4. Hawthorn, Australia: The Council, September 1980. ERIC: ED 194 317.

Bloom, Benjamin S. (Ed.) Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay, 1956.

Bloom, Benjamin S.; Hastings, J. T.; and Madaus, G. F. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill, 1971.

Fielding, Glen D. and Shalock, Del H. Integrating Teaching and Testing: A Handbook for High School Teachers. Monmouth: Oregon State System of Higher Education, January 1985. ERIC: ED 257 821.

Krathwohl, David R.; Bloom, Benjamin S.; and Masia, Bertram B. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook II: Affective Domain. New York: David McKay, 1964.

Mathematical Sciences Education Board. Information Releases. Washington, D.C.: The Board, National Academy of Sciences, 1986.

National Assessment of Educational Progress. Math Objectives: 1985-86 Assessment. Princeton, New Jersey: NAEP, 1985.

National Council of Teachers of Mathematics. An Agenda for Action: Recommendations for School Mathematics of the 1980s. Reston, Virginia: The Council, 1980.

Norton, Mary Ann. Teaching: Improve Your Evaluation Techniques. Arithmetic Teacher. 30: 6-7; May 1983.

Pikaart, Len and Travers, Kenneth J. Teaching Elementary School Mathematics: A Simplified Model. Arithmetic Teacher 20: 332-342; May 1973.

Stiggins, Richard J. and Bridgeford, Nancy J. The Ecology of Classroom Assessment. Journal of Educational Measurement 22: 271-286; Winter 1985.

Swart, William L. Evaluation of Mathematics Instruction in the Elementary Classroom. Arithmetic Teacher 21: 7-11; January 1974.

(General - continued)

Virginia Council of Teachers of Mathematics. Mathematics Assessment for the Classroom Teacher. Charlottesville, Virginia: The Council, 1983.

Wirtz, Robert. The Tyranny of Tests in Elementary School Mathematics. Washington, D.C.: Curriculum Development Associates, April 1979. ERIC: ED 176 950.

Interviews

Callahan, Leroy G. Clinical Evaluation and the Classroom Teacher. February 1973. ERIC: ED 076 640.

Far West Laboratory for Educational Research and Development. Instruments for Individual Assessment of Achievement. Beginning Teacher Evaluation Study Technical Note Series. San Francisco: The Laboratory, September 1975. ERIC: ED 170 304.

Ginsburg, Herbert. The Clinical Interview in Psychological Research on Mathematical Thinking: Aims, Rationales, Techniques. For the Learning of Mathematics 1: 4-11; March 1981.

Hart, Kath. Tell Me What You Are Doing. Mathematics Teaching 99: 32-37; June 1982.

Lankford, Francis G., Jr. What Can a Teacher Learn About a Pupil's Thinking Through Oral Interviews? Arithmetic Teacher 21: 26-32; January 1974.

Reys, Robert E.; Rybolt, James F.; Bestgen, Barbara J.; and Wyatt, J. Wendell. Processes Used by Good Computational Estimators. Journal for Research in Mathematics Education 13: 183-201; May 1982.

Schoen, Harold L. Using the Individual Interview to Assess Mathematics Learning. Arithmetic Teacher 27: 34-37; November 1979.

Weaver, J. Fred. Big Dividends from Little Interviews. Arithmetic Teacher 2: 40-47; April 1955.

Williams, S. Irene and Jones, Chancey O. A Comparison of Interview and Normative Analysis of Mathematics Questions. Princeton, New Jersey: Educational Testing Service, April 1972. ERIC: ED 067 397.

Item Pools

Arter, J. and Estes, G. D. Item Banking for Local Test Development: Practitioners' Handbook. Portland, Oregon: Northwest Regional Educational Laboratory, 1985. ERIC: ED 266 166.

Education Commission of the States. Math Resource Items for Minimal Competency Testing. Denver: The Commission, December 1977. ERIC: ED 173 395.

Fraser, Graham (Ed.). Mathematics Library of Test Items. Sydney, Australia: New South Wales Department of Education, July 1978. ERIC: ED 218 299.

Kahn, Henry F. Needed: An Alternative for Mathematics Textbooks. School Science and Mathematics 79: 476-478; October 1979.

Lieberman, Marcus and others. Behavioral Objectives and Test Items for (1) Primary Mathematics, (2) Intermediate Mathematics, (3) Junior High Mathematics, and (4) High School Mathematics. Downers Grove, Illinois: Institute for Educational Research, 1972. ERIC: ED 066 494, ED 066 495, ED 066 496, ED 066 497.

Instructional Objectives Exchange. Objectives and Test Items for Grades K-9, 10-12. Los Angeles: The Exchange, 1972, 1973, 1981. ERIC: ED 171 768, ED 171 770, ED 171 773, ED 171 785, ED 173 404, ED 173 406.

National Assessment of Educational Progress. Selected Supplemental Mathematics Exercises, National Assessment of Educational Progress. Denver: NAEP, October 1977. ERIC: ED 183 388.

National Assessment of Educational Progress. The Second Assessment of Mathematics, 1977-78, Released Exercise Set. Denver: NAEP, May 1979. ERIC: ED 187 543.

North Carolina. Metrics, The Measure of Your Future: Criterion-Referenced Metrics Tests, Levels K-8. Raleigh: North Carolina State Department of Public Instruction; Winston-Salem: Winston-Salem City Schools, May 1977. ERIC: ED 160 387.

School Mathematics Study Group. Test Batteries, Description and Statistical Properties of Scales -- Kindergarten, Grade 1, Grade 2, Grade 3. ELMA Technical Reports 1-4. Stanford, California: SMSG, 1971. Available on loan from ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Wilson, James W.; Cahen, Leonard S.; and Begle, Edward G. (Eds.). Test Batteries for X-Population, Y-Population, and Z-Population. NLSMA Reports 1-3. Stanford, California: SMSG, 1968. Available on loan from ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Materials Evaluation

Heck, William P.; Johnson, Jerry; Kansky, Robert J.; and Dennis, Dick. Guidelines for Evaluating Computerized Instructional Materials. Reston, Virginia: National Council of Teachers of Mathematics, 1984.

National Council of Teachers of Mathematics. How to Evaluate Mathematics Textbooks. Reston, Virginia: The Council, 1982.

Other Evaluation Techniques

Ash, Michael J. and Sattler, Howard E. A Video Tape Technique for Assessing Behavioral Correlates of Academic Performance. March 1973. ERIC: ED 074 747.

Cornett, J. Alternatives to Paper and Pencil Testing. NASSP Bulletin 66: 44-46; November 1982.

Finstein, Phyllis. Color Their Arithmetic. Arithmetic Teacher 26: 20-22; April 1979.

Greenius, Eric A. Notebook Evaluation Made Easy! Mathematics Teacher 76: 106-107; February 1983.

Hammitt, Helen. Evaluating and Reteaching Slow Learners. Arithmetic Teacher 14: 40-41; January 1967.

Howden, Hilde. An Alternative to Conventional Methods of Evaluation. In The Agenda in Action, edited by Gwen Shufelt. 1983 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1983.

Kissel, Mary Ann and Yeager, John L. An Investigation of the Efficiency of Various Observational Procedures. February 1971. ERIC: ED 048 372.

Lauritzen, Carol. Using Every Pupil Response in Mathematics Education. Arithmetic Teacher 33: 46-47; December 1985.

Noble, John W. Computerized Testing: A New System of Evaluation. Mathematics Teacher 74: 385-388; May 1981.

Peck, Donald M. and Jencks, Stanley M. What the Tests Don't Tell. Arithmetic Teacher 21: 54-56; January 1974.

Reys, Robert E. Testing Mental-Computation Skills. Arithmetic Teacher 33: 14-16; November 1985.

Reys, Robert E. Evaluating Computational Estimation. In Estimation and Mental Computation, edited by Harold L. Schoen. 1986 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1986.

Reys, Robert E. and Bestgen, Barbara J. Teaching and Assessing Computational Estimation Skills. Elementary School Journal 82: 117-127; November 1981.

(Other Evaluation Techniques - continued)

Schminke, Clarence W. The Arithmetic Folder. Arithmetic Teacher 9: 152-154; March 1962.

Wolff, Harry. Oral Testing. Mathematics Teacher 52: 384-387; May 1959.

Problem-Solving Tests

Charles, Randall I. Teaching: Evaluation and Problem Solving. Arithmetic Teacher 30: 6-7, 54; January 1983.

Charles, Randall; Lester, Frank; and O'Daffer, Phares. How to Evaluate Progress in Problem Solving. Reston, Virginia: National Council of Teachers of Mathematics, 1987.

Forsyth, Robert A. and Ansley, Timothy N. The Importance of Computational Skill for Answering Items in a Mathematics Problem-Solving Test: Implications for Construct Validity. Educational and Psychological Measurement 42: 257-263; Spring 1982.

Forsyth, Robert A. and Spratt, Kevin F. Measuring Problem Solving Ability in Mathematics with Multiple-Choice Items: The Effect of Item Format on Selected Item and Test Characteristics. Journal of Educational Measurement 17: 31-43; Spring 1980.

Hofmann, Roseanne. Construction and Validation of a Testing Instrument to Measure Problem Solving Skills. Unpublished Doctoral Dissertation, Temple University, 1986.

Malone, John A.; Douglas, Graham A.; Kissane, Barry V.; and Mortlock, Roland S. Measuring Problem-solving Ability. In Problem Solving in School Mathematics, edited by Stephen Krulik. 1980 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1980.

Mathematical Problem Solving Project. Technical Reports -- II, IV, V. Bloomington, Indiana: Mathematics Education Development Center, Indiana University, 1977. ERIC: ED 168 843, ED 168 846, ED 168 849, ED 168 850, ED 168 851, ED 168 852, ED 168 853.

Romberg, Thomas A. The Development and Validation of a Set of Mathematical Problem-Solving Superitems. Executive Summary of the NIE/ECS Item Development Project. Madison: Wisconsin Center for Education Research, January 1982.

Schoen, Harold L. and Oehmke, Theresa. A New Approach to the Measurement of Problem-solving Skills. In Problem Solving in School Mathematics, edited by Stephen Krulik. 1980 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1980.

Wearne, Diana C. Development of a Test of Mathematical Problem Solving Which Yields a Comprehension, Application, and Problem Solving Score. Unpublished Doctoral Dissertation, University of Wisconsin-Madison, 1976. Dissertation Abstracts International 37A: 6328-6329; April 1977.

Program Evaluation

Alaska State Department of Education. Criteria of Excellence - Mathematics. Juneau: The Department, August 1979. ERIC: ED 215 872.

Buchanan, Aaron D. and Milazzo, Patricia A. Proficiency Verification Systems: A Large-Scale, Flexible-Use Program for Evaluating Achievement in Mathematics. Los Alamitos, California: Southwest Regional Laboratory for Educational Research and Development, April 1977. ERIC: ED 137 369.

Helgeson, Jerry and others. Mathematics Program Assessment and Planning Handbook. Boise: Idaho State Department of Education, 1981. ERIC: ED 213 602.

Hobbs, Eugene and others. Curriculum Review Handbook: Mathematics 1981-82. Oklahoma City: Oklahoma State Department of Education, 1982. ERIC: ED 213 591.

Houser, Larry L. and Heimer, Ralph T. A Model for Evaluating Individualized Mathematics Learning Systems. Arithmetic Teacher 26: 54-55; December 1978.

Maryland State Department of Education. Standards for Successful Mathematics Programs. Baltimore: The Department, 1978. ERIC: ED 162 864.

National Council of Teachers of Mathematics. How to Evaluate Your Mathematics Program. Reston, Virginia: The Council, 1981.

Reidy, Edward F., Jr. and Wallace C., Jr. Skills Achievement Monitoring: Assumptions and Components. May 1980. ERIC: ED 191 657.

Retson, James N. Evaluating Change. In Changing School Mathematics: A Responsive Process, edited by Jack Price and J. D. Gawronski. Professional Reference Series. Reston, Virginia: National Council of Teachers of Mathematics, 1981.

Whitman, Nancy C. Assessing and Improving a School's Mathematics Program, K-8. In The Agenda in Action, edited by Gwen Shufelt. 1983 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1983.

Special Populations

Consuegra, Gerard F. Identifying the Gifted in Science and Mathematics. School Science and Mathematics 82: 183-188; March 1982.

Johnson, Martin L. Identifying and Teaching Mathematically Gifted Elementary School Children. Arithmetic Teacher 30: 25-26, 55-56; January 1983.

Johnson, Orval G. Tests and Measurements in Child Development. San Francisco: Jossey-Bass, 1976. See ERIC: ED 132 207.

Mauser, August J. Assessing the Learning Disabled: Selected Instruments. San Rafael, California: Academic Therapy Publications, 1976. See ERIC: ED 128 438.

Miller, Susan. Tests Used with Exceptional Children: Annotated Bibliography. Des Moines, Iowa: Drake University, July 1975. ERIC: ED 132 773.

Standardized Tests

Alford, Linda E. Alignment of Textbook and Test Content. Arithmetic Teacher 34: 25; November 1986.

Braswell, James and Owens, Douglas T. Mathematics Tests Available in the United States and Canada. Fifth Edition. Reston, Virginia: National Council of Teachers of Mathematics, and Columbus, Ohio: ERIC Clearinghouse for Science, Mathematics, and Environmental Education, 1981.

Buros, Oscar Krisen (Ed.). The Eighth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1978.

Carry, L. Ray. A Critical Assessment of Published Tests for Elementary School Mathematics. Arithmetic Teacher 21: 14-18; January 1974.

Dahle, Mary McMahon. A Procedure for the Measurement of the Content Validity of Standardized Tests in Elementary Mathematics. Unpublished Doctoral Dissertation, University of Southern California, 1970. Dissertation Abstracts International 30A: 5336; June 1970.

Epstein, Marion G. Standardized Tests Can Measure the Right Things. Mathematics Teacher 66: 294, 363-366; April 1973.

Floden, Robert E. and others. Don't They All Measure the Same Thing? Consequences of Selecting Standardized Tests. Research Series No. 25. July 1978. ERIC: ED 167 632.

Freeman, Donald J.; Kuhs, Therese M.; Knappen, Lucy B.; and Porter, Andrew C. A Closer Look at Standardized Tests. Arithmetic Teacher 29: 50-54; March 1982. See also ERIC: ED 179 581, ED 199 047.

Joselyn, E. Gary. An Introduction to Standardized Testing for Teachers and Administrators. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, September 1975. ERIC: ED 177 197.

Lewis, Janice and Hoover, H. D. The Effect on Pupil Performance of Using Hand-Held Calculators on Standardized Mathematics Achievement Tests. April 1981. ERIC: ED 204 152.

McDonald, Margaret. Factors That May Influence Performance on Standardized Tests. In The Agenda in Action, edited by Gwen Shufelt. 1983 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1983.

Mercer, Maryann. The Content of Two Mathematics Achievement Subtests. School Science and Mathematics 78: 669-674; December 1978.

(Standardized Tests - continued)

Mitchell, James V., Jr. (Ed.). The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln, 1985.

Petrosko, Joseph M. and Huano, Linda. An Assessment of the Quality of High School Mathematics Tests. April 1975. ERIC: ED 109 188. See also Journal for Research in Mathematics Education 9: 137-148; March 1978.

Schmidt, William H. Measuring the Content of Instruction. Research Series No. 35. East Lansing: Michigan State University Institute for Research on Teaching, October 1978. ERIC: ED 171 783.

Sheehan, Daniel S. and Davis, Robbie G. The Development and Validation of a Criterion-Referenced Mathematics Battery. School Science and Mathematics 79: 125-132; February 1979.

Wilson, James W. Standardized Tests Very Often Measure the Wrong Things. Mathematics Teacher 66: 295, 367-370; April 1973.

Teacher-Made Tests

Beattie, John and Algozzine, Bob. Testing for Teaching. Arithmetic Teacher 30: 47-51; September 1982.

Beyer, Jim and Sherman, Jon. Why Not a Mathematics Quiz? Arithmetic Teacher 27: 28-29; May 1980.

Brown, F. G. Measuring Classroom Achievement. New York: Holt, Rinehart, & Winston, 1981.

Carlson, Sybil B. Creative Classroom Testing: 10 Designs for Assessment and Instruction. Princeton, New Jersey: Educational Testing Service, 1985.

Epstein, Marion G. Testing in Mathematics: Why? What? How? Arithmetic Teacher 15: 311-319; April 1968.

Evans, S. S.; Evans, W. H.; and Mercer, C. D. Assessment for Instruction. Boston: Allyn & Bacon, 1986.

Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Gronlund, Norman E. Measurement and Evaluation in Teaching. New York: Macmillan, 1984.

Gulliksen, Harold. Creating Better Classroom Tests. Princeton, New Jersey: Educational Testing Service, December 1985. ERIC: ED 268 149.

Halpern, Lynn. A Different Kind of Preparation. Arithmetic Teacher 32: 40-41; October 1984.

Inskeep, James E., Jr. Diagnosing Computational Difficulty in the Classroom. In Developing Computational Skills, edited by Marilyn N. Suydam. 1978 Yearbook. Reston, Virginia: National Council of Teachers of Mathematics, 1978.

Kuhs, Therese; Porter, Andrew; Floden, Robert; Freeman, Donald; Schmidt, William; and Schwille, John. Differences Among Teachers in Their Use of Curriculum-Embedded Tests. Elementary School Journal 86: 141-153; November 1985.

Long, Lynette. Writing an Effective Arithmetic Test. Arithmetic Teacher 29: 16-18; March 1982.

McKillip, William D. Teacher-Made Tests: Development and Use. Arithmetic Teacher 27: 38-43; November 1979.

Miller, Patrick W. and Erickson, Harley E. Teacher-Written Student Tests: A Guide for Planning, Creating, Administrating and Assessing. Washington, D.C.: National Education Association, 1985. ERIC: ED 266 182.